AFRL-RY-WP-TR-2017-0196

ANALYSIS-DRIVEN DESIGN OF REPRESENTATIONS FOR SENSING-ACTION SYSTEMS

Stefano Soatto

University of California, Los Angeles

OCTOBER 2017
Final Report

Approved for public release; distribution is unlimited.

See additional restrictions described on inside pages

STINFO COPY

AIR FORCE RESEARCH LABORATORY
SENSORS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7320
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nationals.

Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RY-WP-TR-2017-0196 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

// Signature//

BERNARD ABAYOWA, Program Manager
Electro-Optic Exploitation Branch

// Signature//

CLARE MIKULA, Branch Chief
Electro-Optic Exploitation Branch

// Signature//

DOUG HAGER, Deputy
Layered Sensing Exploitation Division

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government’s approval or disapproval of its ideas or findings.

*Disseminated copies will show “//Signature//” stamped or typed above the signature blocks.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YY)*	**2. REPORT TYPE**	**3. DATES COVERED *(From - To)***
October 2017	Final	26 September 2011 – 30 June 2017

4. TITLE AND SUBTITLE ANALYSIS-DRIVEN DESIGN OF REPRESENTATIONS FOR SENSING-ACTION SYSTEMS	**5a. CONTRACT NUMBER** FA8650-11-1-7156
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 61101E/62204F
6. AUTHOR(S) Stefano Soatto	**5d. PROJECT NUMBER** 1000/6095
	5e. TASK NUMBER 11
	5f. WORK UNIT NUMBER Y0PH
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of California, Los Angeles 420 Westwood Plaza Los Angeles, CA 90095-1596	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory Sensors Directorate Wright-Patterson Air Force Base, OH 45433-7320 Air Force Materiel Command United States Air Force Defense Advanced Research Projects AgencyDARPA/DSO 675 North Randolph Street Arlington, VA 22203	**10. SPONSORING/MONITORING AGENCY ACRONYM(S)** AFRL/RYAT **11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)** AFRL-RY-WP-TR-2017-0196

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This material is based on research sponsored by Air Force Research laboratory (AFRL) and the Defense Advanced Research Agency (DARPA) under agreement number FA8650-11-1-7156. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation herein. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies of endorsements, either expressed or implied, of Air Force Research Laboratory (AFRL) and the Defense Advanced Research Agency (DARPA) or the U.S. Government. Report contains color.

14. ABSTRACT

We have developed what is, to the best of our knowledge, the first complete theory of representation for decision and control task, which has shown not only to encompass and explain all known phenomenology in deep neural network-based representation learning, but also to predict phenomena that were thus far unexplained.

15. SUBJECT TERMS

actionable information, representations, multiview

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Bernard Abayowa
Unclassified	Unclassified	Unclassified	SAR	9	**19b. TELEPHONE NUMBER** *(Include Area Code)* N/A

Table of Contents

1.0 Scope

The project's initial aim was to develop analytical and computational tools to actively infer information-driven representations of natural and man-made scenes from visual sensory data streams. As of the date of completion of the project, we have developed what is, to the best of our knowledge, the first complete theory of representation for decision and control task, which has shown not only to encompass and explain all known phenomenology in deep neural network-based representation learning, but also to predict phenomena that were thus far unexplained.

This project dovetailed the Phase I MSEE (Mathematics of Sensing, Exploitation and Execution) effort that started before the explosion in popularity of Deep Learning. It should be remarked that the foundations of the theory not only withstood this trend, but naturally served to provide theoretical foundations to it.

2.0 Research Accomplishments

In this section we report our accomplishments and breakthroughs in chronological order, leading to the most recent development of a complete theory. For a description of accomplishment organized according to tasks set forth in the original proposal, please refer to interim reports.

2.1 Actionable Information

At the beginning of the project, in 2015, we had formulated the seeds of a theory of information in support of decision and control (as opposed to transmission and storage) tasks. At the heart of this was the definition of a notion of information (named Actionable Information) that discounted nuisance variability in the data. While for traditional signal processing such nuisances were limited to (typically white, zero-mean, additive) noise, for the case of images they include occlusion, scaling, viewpoint-induced deformations and illumination changes. Collectively, such sources of variability are infinitely more complex that the intra-class variability at the heart of the decision or control task.

For the sake of example, consider the "Fetch scenario" used to guide development of a project: At each instant of time, the system has to make a decision as to whether or not the object of interest is present in the scene (1 bit), and if so take a control action (2 to 6 degrees of freedom) to move towards it. The data (streaming images) are high-dimensional (millions of pixels per frame, or tens of megabytes per second), and more importantly highly variable in a manner that is irrelevant to the task at hand: We want to be able to make the decision about the presence of the object regardless of where it is (viewpoint), how much of it is visible (occlusion), irrespective of the reflectance properties of nearby objects, and of the illumination of the scene. Although the variability in the images is vast (and therefore the "information" needed to store them and transmit them), the actionable component is small. At that time, there was no notion of information that allowed capturing this fundamental phenomenon, that is particularly severe for imaging data where most of the variability can be ascribed to nuisance factors.

2.2 Optimal Representations

In 2015, we formalized the defining characteristics of a representation, which ideally would be the function of the data that contains all and only the Actionable Information (AI). However, rather than starting from AI, we started from basic principles of statistical decision and information theory:

Sufficiency, minimality, invariance. In (Soatto & Chiuso, 2016) we published our findings at ICLR, the International Conference on Learning Representations. We defined a representation as a **sufficient invariant** statistic. We then showed that, if such a statistic could be found, it would maximize Actionable Information. At that time, however, we had no knowledge on how to compute such sufficient invariants.

2.3 Contrast with Other Nascent Theories

It should be noted that already the definitions above contrast with other developing theories, including those championed by Stephane Mallat, Tomaso Poggio, Richard Baraniuk, and Yoshua Bengio. First, as we were doing prior to 2011, all of them are focused on **maximal invariance**, rather than **sufficient invariance**. Maximal invariance, often referred to improperly as "selectivity", refers to the ability of reconstructing a copy of the data from the representation, under the action of a (group) invertible nuisance. For instance, the theory of Mallat aims to compute a "transform" from which the original data can be reconstructed. But the goal is not to reconstruct the data, but to use it for decision and control. Indeed, for the purpose of correspondence, the Scattering Transform performs even worse than SIFT, which it provably generalizes! Similarly, the work of Poggio and co-workers is only valid for locally compact groups. The problem is that, in vision, occlusions are not a group, so any theory that hinges on nuisances having a group structure will not withstand reality. Indeed, their theory can be seen as a special (and degenerate) case of ours. Finally, the work of Bengio is generally based on vague claims that are not followed through by analysis, but it is generally focused on "disentanglement" of the components of a representation, which is however never formalized, computed, or bounded in any elements of the (vague and verbose) work of Bengio.

2.4 Engineered Features

We simplified the framework to the simplest possible representational problem in visual processing: Correspondence under similarity. The task is to decide whether two images portray the same scene (binary classification). The model is that, if there is correspondence, the two images are locally related by a similarity transformation of the domain (induced by viewpoint changes away from occlusions) and a similarity transformation of the rage (induced by local changes of illumination away from cast shadows). In this case, we showed that a sufficient invariant can be computed in closed form. In (Dong and Soatto, 2015), we showed that the resulting local sufficient invariant (a "descriptor", or "feature") was very similar to SIFT/HOG – the most popular choice of descriptors then – but for the fact that the latter were not marginalizing scale. Adding scale marginalization (few lines of code) improved performance of the resulting descriptor (called DSP-SIFT, for domain-size pooling SIFT) by considerable margins on public benchmarks (up to 30% improvements in mean-average precision on the Mikolajczyk dataset). This was, however, the most trivial learning problem, consisting of a single training image, and a test image, with a binary classification task and an explicit nuisance model. It was interesting that the resulting descriptor closely resembled the first layer of a convolutional neural network, that at that time was gaining traction as the method of choice even for binary correspondence (although that that time DSP-SIFT handily beat deep neural networks, as well as Mallat's Scattering Transform despite its theoretical guarantees).

2.5 Localization Tasks

What is a nuisance depends on the task: For object detection or recognition, vantage point is a nuisance as we want to be able to detect an object regardless of where it is. However, for localization vantage point is a goal, whereas the reflectance properties of object are irrelevant. Since localization is one of the tasks underlying most control applications and interactions with physical scenes, we devoted a considerable amount of effort to this task, which led to the first complete analysis of the observability and identifiability of visual-inertial sensor fusion, and a real-time implementation of a robust visual-inertial odometry and mapping system that still performs at state-of-the-art levels despite competition from private companies, including Google/Tango (that hired one of the students supported on the project, and since then incorporated some of the published ideas into their products). This includes (Fei et al., 2016) that developed efficient search methods for large-scale mapping and loop closure. In addition to ego-localization, localizing objects in scene is one of the central problems of visual perception, and we made significant progress with (Taylor et al., 2015), where we were able to causally segment objects using persistence of visibility. Both students involved in that project joined autonomous driving startups: Uber (B. Taylor), and Zoox (V. Karasev).

2.6 A Leap in the Theory

Up to 2015, we subscribed to the notion that, since the AI is thin in the (imaging) data, an optimal representation should be far "smaller" than the data. In line with work on Sufficient Dimensionality Reduction (work of M. Jordan, D. Cook and co-workers), we were attempting to construct statistics (deterministic functions of the data) that had lower dimension, i.e. fewer free parameter than the dimension of the data. However, "small" needs not be measured in terms of dimension (number of parameters). Instead, it can be measured by the amount of information such parameters contain (which is upper-bounded by the dimension, but could be far lower). This led to the notion of Information Dropout (Achille and Soatto, 2016), that turned out to be very closely related to methods that practitioners were already using in training deep neural networks.

This realization spawned a new direction in the theory, that now aimed at construction high-dimensional, low-information representations. It should be noted that such a theory ended up being perfectly consistent with both the initial theory of Actionable Information, as well as with the practice of Deep Learning. It should also be noted that this theory is closely connected with the Information Bottleneck Principle, developed by Naftali Tishby and coworkers, that was the starting point of our investigators in the MSEE program leading into this project, well before deep neural networks were popularized.

3.0 Unforeseen Accomplishments

The central points of the newly developed theory are as follows (i) the notion of sufficiency and minimality are captured by the Information Bottleneck Lagrangian. This was well known, but it had no known relation to invariance. In addition, until recently it was not known how to compute, let alone optimize, the IB Lagrangian; we did so for a very general class of (deep neural network) models; (ii) we showed that for a sufficient statistic, minimality guarantees invariance to nuisance factors; (iii) we showed that the IB Lagrangian can be re-written as a data term that is precisely the cross-entropy loss most commonly used for training deep neural networks, and a regularizer that is not ordinarily considered; (iv) we showed that the coefficient that modulates this regularizer relates to the amount of information a network is allowed to store in the weights, and predicted a sharp phase transition between overfitting and underfitting of random labels, observed in practice. Finally, (v) we showed that, if one could write an explicit loss function to minimize overfittig, the minimizer would be dual to that obtained by minimizing an Information Bottleneck Lagrangian.

4.0 Future Work

We are currently in the process of exploring connections that the theory has with PAC-Bayes (Probably Approximately Correct Bayesian) theories, as well as with Kolmogorov Complexity (algorithmic information theory). We are also exploring algorithmic implications of the theory, including the role of so-called “flat minima” found in stochastic gradient descent play in minimizing generalization error.

5.0 Partial Misses

Although we consider the outcome of the project overall positive and beyond our expectation at the outset, we accomplished less than expected on the development of the experimental testbed we had conceived of. In particular, after the initial collaboration to port Corvis (our visual-inertial framework) into the R-HEX robot developed in the lab of Prof. Dan Koditscheck, we have not followed up that development.

A preliminary demonstration of the concepts developed has been made using the FETCH scenario. That consisted of a toy quadrotor connected to a laptop computer via a radio-link, with video processed live on the laptop and control and planning commands sent to the quadrotor. The task was to (a) learn the appearance of a single object, (b) find the object in an unknown scene, which requires (c) localizing the platform relative to the scene and (d) planning a trajectory that would decrease the uncertainty of the location of the object.

We have relaxed some of the initial assumptions made in the FETCH demo and allowed generic shape of the environment. We have also isolated the sensory system and demonstrated, live at CVPR 2016, the visual-inertial-semantic perception system, the first of its kind.

We did not close the loop. Part of this is due to the fairly high-effort, low-intellectual-yield nature of the work entailed, so we preferred focusing on the theoretical aspects of the development. Part is due to the absence of potential partners for robotics at UCLA, where robotics is absent in the Computer Science and Electrical Engineering departments, and nascent (and under-represented) in the Mechanical Engineering department. We did develop a turtlebot testbed, and we were able to replicate the results of the FETCH demo with a more stable platform (one that can operate for more than 15’ without the need to recharge), and the result has been summarized in the Master Thesis of Isaac Deutsch at the ETH-Zurich.

All in all, we feel our project expanded well beyond the envelope foreseen in the initial proposal, in some cases at the expenses of the more experimental component of the project as initially conceived.

6.0 Personnel and Organization

The team consists of a single investigator (Stefano Soatto) and partial support for students totaling a one-student year. Partially supported students will include any among Konstantine Tsotsos, Xiaohan Fei and Pratik Chaudhari. Please refer to the financial closing package for details on personnel and involvement.

7.0 References

X. Fei, K. Tsotsos and S. Soatto. A simple hierarchical averaging data structure for loop closure, *Proceeding of the European Conference on Computer Vision (ECCV)*, 2016.

S. Soatto and A. Chiuso. Visual scene representations: sufficiency, minimality, invariance and deep approximations, *Proceeding of the International Conference Learning and Representation (ICLR)*, 2016.

S. Soatto, J. Dong and N. Karianakis. Visual scene representations: Scaling and occlusion in convolutional architectures, *Proceeding of the International Conference Learning and Representation (ICLR) Workshop*, 2015.

B. Taylor, V. Karasev and S. Soatto. Causal video object segmentation from persistence of occlusions, *Proceeding of Computer Vision and Pattern Recognition* (CVPR), 2015.

J. Dong and S. Soatto. Domain size pooling in local descriptors: DSP-SIFT, *Proceeding of Computer Vision Pattern Recognition (CVPR)*, 2015.

J. Dong, J. Hernandez, J. Balzer, D. Davis and S. Soatto. Multiview feature engineering and learning, *Proceeding of the Computer Vision and Pattern Recognition (CVPR)*, 2015.

D. Davis. Asymmetric sparse kernel approximations for nearest neighbor search, *Proceeding of Computer Vision and Pattern Recognition (CVPR)*, 2013.

V. Karasev, A. Ravichandran and S. Soatto. Active frame, location, and detector selection for automated and manual video annotation, *Proceeding of Computer Vision and Pattern Recognition (CVPR)*, 2013.

A. Achille and S. Soatto. Information Dropout: Learning optimal representations through noisy computation, submitted to *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 2016.